ASSIGNMENT 4

Introduction to Data Science

Presented to: Dr Muhammad Sharjeel

Noor Zehra
Sp20-BCS-123
Sp20-bcs-123@cuilahore.edu.pk

Q1: Provide responses to the following questions about the dataset.

1. **How many instances does the dataset contain?**
   Answer: 80 instances (initially, 85 after question 4)

2. **How many input attributes does the dataset contain?**
   Answer: 7 (height, weight, shoe_size, beard, hair_length, scarf, eye_color)

3. **How many possible values does the output attribute have?**
   Answer: 2 (Male or Female)

4. **How many input attributes are categorical?**
   Answer: 4 (Beard, hair_length, scarf, eye_color)

5. **What is the class ratio (male vs female) in the dataset?**
   Answer: **23**: **17** (Males =46 , Females=34)

Q2: Apply Random Forest, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with standard train/test split ratio and answer the following questions.

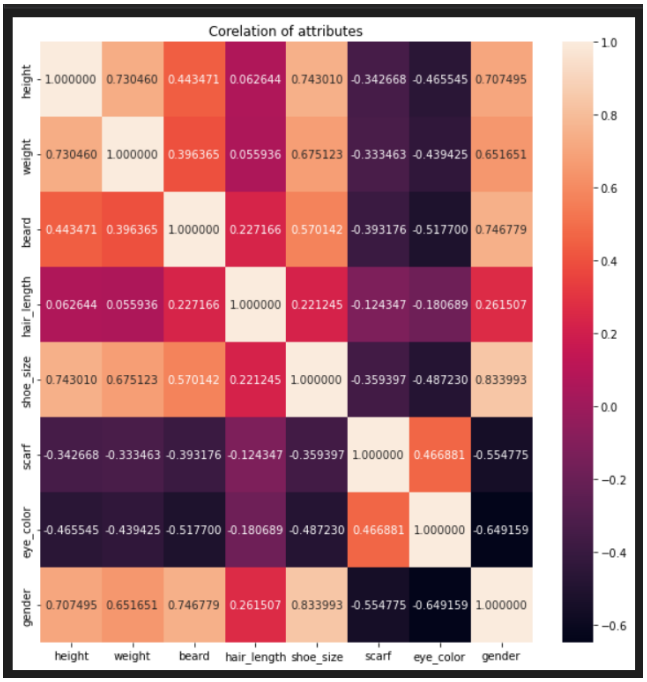
1. **How many instances are incorrectly classified?**

	Random Forest	Support Vector Machines	Multilayer Perceptron
False Positive	0	0	0
False Negative	0	1	1
Total incorrectly identified	0	1	1

2. **Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.**
   The accuracy increased as the number of the incorrectly identified cases decreased. There were 0 wrongly classified instances, using all the algorithm classification.

	Random Forest	Support Vector Machines	Multilayer Perceptron
False Positive	0	0	1
False Negative	0	0	0
Total incorrectly identified	0	0	1

3. Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?
Answer: Two most powerful attributes in the prediction task are 'shoe size' and 'beard'. We can see that after visualizing the data and making a heatmap that shows correlation of features with gender. The correlation factor of shoe_size and beard are most positive. Logically thinking too, beard is quite distinctive feature between men and women.



4. Try to exclude these 2 attribute(s) from the dataset. Rerun the experiment (using 80/20 train/test split), did you find any change in the results? Explain.

The accuracy dropped. There were wrongly classified instances. Although, the number of wrongly classified instances wasn't large due to data set being small.

	Random Forest	Support Vector Machines	Multilayer Perceptron
False Positive	0	0	0
False Negative	1	0	1
Total incorrectly identified	1	0	1

Q3: Apply Decision Tree Classifier classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F1 score for both cross-validation strategies. Note: You are free to choose any parameter values for both cross-validation strategies, however, you have to provide these values in your submission document.

P-Out value used is 2.
For Monte Carlo n is 5.

Q4: Add 5 sample instances into the dataset (you can ask your friends/relatives/sibling for the data). Rerun the ML experiment (using Python) by training the model using Gaussian Naïve Bayes classification algorithm and all the instances from the gender prediction dataset. Evaluate the trained model using the newly added test instances. Report accuracy, precision, and recall scores. Note: You have to add the test instances in your assignment submission document.

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
70	176	yes	short	44	no	black	male
78	193	yes	bald	41	no	black	male
72	182	no	medium	43	no	blue	male
65	158	no	long	38	yes	brown	female
68	160	no	medium	39	no	black	female